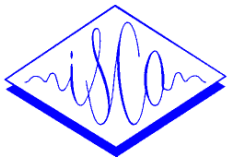
ISCA Archive
http://www.isca-speech.org/archive



ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

RECOGNITION OF SPONTANEOUSLY PRONOUNCED TV ICE-HOCKEY COMMENTARY

Josef Psutka, J. V. Psutka, Pavel Ircing, and Jan Hoidekr

Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic
{psutka, psutka_j, ircing, hojan@kky.zcu.cz}

ABSTRACT

This paper describes our effort with an automatic transcription of TV ice-hockey commentaries[1]. The ice-hockey matches were played during the World Championships 2000 and 2001 in St. Petersburg (Russia) and Hannover (Germany), respectively and were transmitted by the Czech TV channels NOVA and CTV1 with an accompanying commentary of Robert Záruba. Annotation rules designed for the processing of a commentary comprising specific background noise are formed and a list of non-speech events is proposed. A baseline ASR system was built using commentaries of 15 training matches and tested using utterances randomly selected from 4 tested matches. Several types of adaptive lexicons were designed to decrease a number of OOV words and improve recognition accuracy.

1. INTRODUCTION

Automatic transcription of a TV commentary accompanying an ice-hockey match is usually a hard task due to the spontaneous speech of a commentator put often into a very loud background noise created by the public, music, siren, drums, whistle, etc. Although a vocabulary of an ice-hockey commentary seems to be relatively well limited we show that the lexicon formed by collecting words of 15 ice-hockey matches does not cover the commentary of new matches sufficiently and a high OOV (Out Of Vocabulary) rate brings about many recognition problems. It is probably caused by a high degree of inflection, a high degree of derivations (prefixes and suffixes) and the relatively free word order of the Czech language as a representative of the large family of Slavic languages [1]. We try to alleviate the high OOV rate by a suitable knowledge-based selection of recognition lexicon. This approach is based on the usually known list of players' names that can be added to the lexicon in all expected grammatical cases before the match starts. Results of several experiments with an adaptive lexicon showed that we are able to decrease the OOV rate by 3 to 4% simultaneous- ly both with decreasing the number of words in the lexicon by 7 to 12% and improving recognition accuracy by 6 to 13%. All recognition experiments were performed for zero-, uni-, and bi-gram language models (LMs) proposed by the Good-Turing discounting technique. Let us mention that this paper does not try to solve problems connected with the reduction of often very loud and non-stationary noise in the background of the commentary of most matches.

[1] This work was funded by the Ministry of Education of the Czech Republic, project No. MSM234200004, and project No. LN00A063.

This paper is organized as follows. Section 2 describes the way in which the ice-hockey speech data was acquired and gives the rules for an annotation process including the list of non-speech events. Then, Section 3 gives some lexical statistics on collected corpora and Section 4 depicts selected recognition experiments using adaptive vocabulary. Finally, Section 5 gives short conclusions with an outline of future work.

2. DATA ACQUISITION AND SPEECH CORPORA ANNOTATION

Data for this task was collected during the ice-hockey World Championships 2000 held in St. Petersburg and 2001 held in Hannover, respectively. Several tens of matches transmitted by the Czech TV channels NOVA and CTV1 were received in our lab by the PC equipped with a common TV card. Only audio stream with the commentary was recorded at 44.1 kHz with 16-bit resolution. The majority of these matches was accompanied by the commentary of Robert Záruba, one of the most experienced Czech ice-hockey commentators. For our initial experiments we decided to build the ASR system adapted only to the voice of this commentator. So we selected 19 matches; 15 of these were specified for the training of acoustic and language models including the construction of an optimum lexicon. Remaining 4 matches created a test material.

Speech data was annotated using special annotation software Transcriber 1.4.1. It allows manually segmenting, labeling and transcribing speech signal for later use in automatic speech processing. Transcriber is freely available from the Linguistic Data Consortium (LDC) web site http://www.ldc.upenn.edu/. The rules proposed for an annotation process are as follows:

- Audio files were segmented into parts, which roughly correspond to sentences.
- Each segment was labeled with marks giving the beginning of the segment <time ti> <s skp#>, where the ti is the time measured in seconds and the spk# is the speaker ID determined according to the following items:
  Spk1 … commentator (Záruba)
  Spk2 … co-commentator
- All numbers were transcribed as words.
- Sometimes the commentator said words in another language than Czech. They were especially names of players and names of teams. Such words were written in their original language and the Czech pronunciation was enclosed in []:

  New[ňů] York[jork] Islanders[ajlendrs]

- If the sentence was spoken incompletely due to the truncation, the spoken fragment of the sentence was delimited by „~“, e.g. “Čechmánek dostal ~“

- Non-speech sounds like laughter, coughing and loud breath were transcribed as _cg_, _la_ and _lb_, respectively.
- Other disfluencies in speech were marked with one of the following marks: _uh_, _um_ or _er_.
- If the commentary was completely unintelligible, the transcription was _ui_.
- Non-speech sounds created by a noisy environment like music, public, siren, drum and other noise were transcribed as _mu_, _sr_, _dr_ and _ns_.
- If the non-speech event overlapped a spoken lexical item, the descriptor was placed close to the item that was overlapped. Characters „<“ or „>“ were used when the non-speech event was co-occurred before or after lexical item. When non-speech event occurred for more than one lexical item, there were used characters „/“ and „\“ for the beginning and the end of a non-speech event, respectively.

The complete list of non-speech events used for the annotation of a commentary of ice-hockey matches is given in Tab 1.

ER	_er_	LOUD BREATH	_lb_
UH	_uh_	COUGHING	_cg_
UM	_um_	LAUGHTER	_la_
MUSIC	_mu_	BACKGROUND_MUSIC	_bm_
PUBLIC	_pb_	BACKGROUND_VOICE	_bv_
SIREN	_sr_	SIREN_PUBLIC	_sp_
WHISTLE	_wh_	WHISTLE_PUBLIC	_wp_
DRUM	_dr_	DRUM_PUBLIC	_dp_
NOISE	_ns_	UNINTELLIGIBLE	_ui_

Table 1: Complete list of all non-speech events

An example of a short segment of an annotated commentary of the match between teams of the Czech Republic and Canada is as follows:

. . .
<t 323.451> <s spk1> _si_> Dopita zasáhl kotouč vysokou hokejkou.
<t 326.361> <s spk1> _pb_/ Zdá se mi, že si Kanaďané /_pb_ _dp_/ nechávají některé zbraně ukryty v dnešním zápase, velice málo hraje například /_dp_ _pb_/ Steve [stýv] Sullivan [saliven], který měl v Chicagu [šikágu] výtečnou sezónu /_pb_.
. . .

3. LEXICAL STATISTICS OF COLLECTED CORPORA

The first part of collected corpora consists of transcribed commentaries of 15 matches. To better build the final recognition lexicon we were interested in the following characteristics:

- a number of **To**kens uttered in the **M**atch (in Table 2 these characteristics are shown in the column depicted **To_M**)
- a number of different **Wo**rds in the **M**atch (**Wo_M**)
- a cumulative number of **Wo**rds in the **V**ocabulary including the given match (**Wo_V**)
- a number of new words in the match (compared with the actual cumulative vocabulary) (**NWo_M**)
- a **Le**ngth of the **M**atch measured in minutes (**Le_M**)
- a **Sp**eech **Ra**te of the commentator in the match (**Sp_Ra**).

All statistics are depicted in Table 2 including several cumulative numbers showed in the last row. Looking at the numbers in Table 2 we can see that we collected more than one hundred thousand tokens which seems to be a sufficient number for building *n*-gram-based LMs of a baseline ASR system. A relatively high number of unseen words encountered in the new matches (the column **NWo_M)** was caused partly by a portion of new names (pronounced in various grammatical cases), and partly by new words or the words already seen but now with different endings (in different grammatical cases). It is now not quite evident and very hard to predict how many further matches should be collected to reach only several tens of unseen words in a new commentary (compared with an actual lexicon). The realistic estimate is twenty or thirty but the behavior of a related characteristic of the OOV rate could be also considered.

Match	**To_M**	**Wo_M**	**Wo_V**	**NWo_M**	**Le_M** [min]	**Sp_Ra** [w/min]
1	9,096	2,509	2,509	2,509	116.3	78.2
2	6,575	2,188	3,711	1,202	106.9	61.5
3	7,562	2,268	4,794	1,083	119.5	63.3
4	6,544	2,220	5,761	967	104.5	62.6
5	6,897	2,127	6,531	770	108.0	63.9
6	6,383	2,024	7,195	664	105.4	60.6
7	7,463	2,061	7,826	631	107.3	68.7
8	5,498	1,859	8,337	511	106.3	71.2
9	7,268	2,252	8,957	620	109.8	50.1
10	6,457	2,145	9,511	554	101.0	72.0
11	8,191	2,218	10,246	735	109.5	59.0
12	6,829	2,002	10,745	499	101.9	67.1
13	5,643	1,766	11,231	486	98.2	57.5
14	7,862	2,357	11,949	718	103.9	75.7
15	7,602	2,104	12,359	410	104.9	72.5
∑	105,149	-	12,359	-	1,603.1	Av=65.6

Table 2: Statistics on commentaries of the training matches

A relatively low level appears to be for the speech rate (numbers in the column depicted **Sp_Ra**, with an average value 65.5 words/minute). However, analyzing commentaries we can find out time segments with a very high speech rate and parts without a speech commentary, because a video stream sometimes provides enough information for TV spectators. Similar statistics, as for 15 training matches, were also computed for 4 test matches, see Tab 3.

Match	**To_M**	**Wo_M**	**Le_M** [min]	**Sp_Ra** [w/min]
Test 1	7,394	2,071	99.1	74.6
Test 2	7,134	2,107	105.4	67.7
Test 3	5,925	1,946	100.3	59.1
Test 4	7,370	2,214	105.5	69.9

Table 3: Statistics on commentaries of the test matches

Our further interest was to explore how the lexicon of the training matches covers commentaries of four test matches. We successively compared the cumulative lexicons (lexicons arisen collecting words of the first *n* training matches) with a sequence of words (more exact tokens) present in particular test matches. So we computed the corresponding number of OOV (Out Of

Vocabulary) words and also the OOV rate, which is the OOV divided by the total number of tokens in the given test match (column **To_M**, in Table 3). Very interesting sequences of the OOV and/or the OOV rate (OOV_r) for individual test matches are enumerated in Tab 4 and the dependencies of the OOV rate for increasing number of matches are shown in Fig 1.

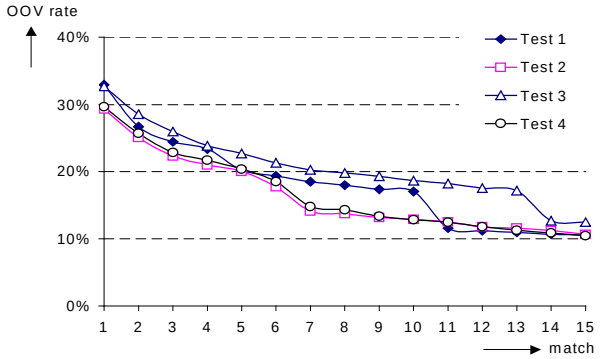


Figure 1: Dependency of the OOV rate on the increasing number of training matches

It is evident from the results given in Tab 4 and depicted in Figure 1, that a level of the OOV rate (greater than 10%) is too high to obtain a reasonable ASR system (every tenth word is not in the lexicon). This problem could be solved by collecting much more speech data or by creating knowledge-based adaptive lexicon.

Match	Test 1		Test 2		Test 3		Test 4	
	OOV	OOV_r [%]	OOV	OOV_r [%]	OOV	OOV_r [%]	OOV	OOV_r [%]
1	2,435	32.9	2,090	29.3	1,938	32.7	2,185	29.7
2	1,974	26.7	1,790	25.1	1,692	28.6	1,895	25.7
3	1,808	24.5	1,594	22.3	1,539	26.0	1,684	22.9
4	1,729	23.4	1,497	21.0	1,414	23.9	1,599	21.7
5	1,497	20.3	1,430	20.0	1,349	22.7	1,501	20.4
6	1,432	19,4	1,264	17,7	1,262	21.3	1,365	18.5
7	1,367	18.5	1,007	14.1	1,201	20.3	1,092	14.8
8	1,330	18.0	979	13.7	1,173	19.8	1,055	14.3
9	1,285	17.4	942	13.2	1,145	19.3	983	13.3
10	1,260	17.0	919	12.9	1,107	18.7	947	12.9
11	857	11.6	892	12.5	1,080	18.2	919	12.5
12	828	11.2	841	11.8	1,040	17.6	869	11.8
13	810	11.0	828	11.6	1,017	17.2	831	11.3
14	790	10.7	802	11.2	753	12.7	801	10.9
15	779	10.5	760	10.7	740	12.5	772	10.5

Table 4: Dependency of the OOV and OOV rate measured on the test matches using successively created "training" lexicon

4. RECOGNITION EXPERIMENTS WITH ADAPTIVE LEXICON

The baseline ASR system was designed to verify the feasibility of this task. The data for training both acoustic and language models was taken from all 15 training matches. The MFCC parameterization including delta and delta-delta coefficients was used. The basic speech unit of the ASR system was a triphone. Each individual triphone was represented by three states HMM with a continuous output probability density function assigned to each state. 8 mixtures of multivariate Gaussians were used for each state. As the number of Czech triphones is too large, phonetic decision trees were used to tie states of Czech triphones. No noise reduction method was applied both on the speech signal side and the acoustic model side. Recognition experiments were performed with zero-, uni-, and bi-gram LMs proposed by the Good-Turing discounting technique. The hidden Markov model toolkit HTK was used for the acoustic modeling and the SRILM toolkit [2] for the language modeling. All experiments were performed using both the HTK toolkit and the AT&T decoder [3].

To better appreciate a quality of designed LMs we enumerated the perplexity *PPL* of the test corpora using LMs built on the manual transcriptions of 15 training matches. Since it is known that $PPL=2^{LP}$ and as we computed the perplexity also using unseen words (OOV words) we had to modify the well-known formula for the computation of the "log probability" *LP* as follows

$$LP = -\frac{1}{S-\mathrm{OOV}} \sum_{i=1}^{S} \log_2 P(w_i \,|\, \Phi(w_{i-1} \ldots w_1)),$$

where $\Phi(w_{i-1} \ldots w_1)$ is a suitable function of a "word history", and S is the number of word tokens in the given test match (which corresponds to the **To_M,** see Table 3). Let us mention that both $P(w_i \,|\, \Phi(w_{i-1} \ldots w_1))$ and $\log_2 P(w_i \,|\, \Phi(w_{i-1} \ldots w_1))$ were considered to be equal zero if w_i belonged to the group of OOV words.

In Tables 5, 6, 7, and 8 you can find results of 4 series of recognition experiments. From each test match we randomly selected 100 utterances which created the test material. In the head of each table you can see (for each test match) the number of words in the "active" lexicon, and the OOV rate (OOV_r), which is here computed for the whole test match (not only for 100 selected test utterances). Similarly the perplexity *PPL* computed using zero-, uni-, bi-, and tri-gram statistics holds for the whole test match. The accuracy (*Acc*) of recognition experiments was determined for 100 selected utterances and for zero-, uni-, and bi-gram LMs. We did not perform recognition experiments for trigrams because we suspect that there is not enough training data to build a reliable trigram LM.

- **The full lexicon.** In these experiments we used the "original" lexicon collected from commentaries of 15 training matches.

LM	Test 1 12,359 [words] 10.5 [OOV_r]		Test 2 12,359 [words] 10.7 [OOV_r]		Test 3 12,359 [words] 12.5 [OOV_r]		Test 4 12,359 [words] 10.5 [OOV_r]	
	PPL	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]
Zero-	37k3	38.1	38k2	38.8	47k5	14.0	37k3	43.2
uni-	1k8	52.9	1k9	61.8	2k0	23.9	1k8	52.4
bi-	629	63.5	560	68.1	645	33.0	543	62.5
tri-	594	-	527	-	617	-	526	-

Table 5: Results of the first series of recognition experiments

Examining the sets of OOV words and also the corresponding recognition results for particular test matches (given in Table 5) we found out that a large portion of OOV

words, which causes difficulties during recognition, usually consists of new names of players, who were active in the given match. On the other hand, there are a lot of names accumulated in the lexicon, which are "dead", that means, which were uttered only several times and there is a very low probability that these words will be pronounced in some of next commentaries. Because the list of players is usually known before a match starts we can adapt the actual lexicon by adding new names in all practicable grammatical cases and/or we can remove the infrequent names from the lexicon and to work only with "living" names. If we remove and/or add some words to the lexicon we have to modify the statistics of the LMs. In our approach we used a LM with a class-based approach to the group of names, which makes a modification of LM statistics easily possible. The designed knowledge-based approach with an adaptive lexicon was evaluated on the following set of experiments:

- **The full lexicon complemented by the list of players active in the given match.** These experiments were performed with the original lexicon (12,359 words) complemented by the list of names of players active in the given match (names were added in all applicable grammatical cases). Results of this series of experiments can be seen in Tab 6.

LM	**Test 1** 12,442 [words] 6.9 [OOV_r]		**Test 2** 12,461 [words] 5.7 [OOV_r]		**Test 3** 12,513 [words] 7.2 [OOV_r]		**Test 4** 12,418 [words] 6.7 [OOV_r]	
	PPL	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]
zero-	25k0	50.0	22k0	45.1	26k0	20.7	24k4	47.5
uni-	2k1	69.6	2k5	68.5	2k1	37.4	2k2	62.7
bi-	725	76.5	782	74.2	725	45.7	755	69.1
tri-	663	-	708	-	675	-	718	-

Table 6: Results of the second series of recognition experiments

- **The full lexicon without all names but complemented by the list of players active in the given match.** These experiments were performed with the original lexicon from which all names were removed and which was complemented by the list of names of players active in the given match (names were added in all applicable grammatical cases). Results of this series of experiments can be seen in Tab 7.

LM	**Test 1** 10,816 [words] 7.3 [OOV_r]		**Test 2** 10,850 [words] 6.6 [OOV_r]		**Test 3** 10,861 [words] 8.2 [OOV_r]		**Test 4** 10,801 [words] 8.2 [OOV_r]	
	PPL	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]
zero-	22k5	50.9	20k9	46.0	24k9	24.2	24k8	46.8
uni-	1k6	70.9	1k8	69.4	1k7	39.1	1k6	61.9
bi-	548	77.0	580	74.3	574	50.0	532	68.5
tri-	502	-	527	-	535	-	506	-

Table 7: Results of the third series of recognition experiments

The full lexicon without all names but complemented by the list of players active in the given match and the lexicon of most frequent names. These experiments were performed with the original lexicon from which all names were removed and which was complemented on the one hand by the list of names of players active in the given test match (same as in the last series of experiments) and on the other hand by the names of players that appeared in the original lexicon more than three times. Results of experiments are showed in Tab 8.

LM	**Test 1** 11,368 [words] 7.1 [OOV_r]		**Test 2** 11,370 [words] 6.2 [OOV_r]		**Test 3** 11,435 [words] 7.8 [OOV_r]		**Test 4** 11,320 [words] 7.5 [OOV_r]	
	PPL	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]	*PPL*	*Acc* [%]
Zero-	23k2	50.0	21k1	46.7	25k2	22.5	24k1	48.5
uni-	1k8	69.2	2k1	68.3	1k9	37.8	1k9	62.9
bi-	638	75.7	674	74.0	649	45.7	641	69.3
tri-	584	-	611	-	605	-	610	-

Table 8: Results of the fourth series of recognition experiments

Analyzing recognition results we can confirm a very positive influence of the knowledge-based adaptive lexicons on the level of the OOV words, which decreased by 3 to 4% simultaneously with a dramatic increase in the recognition accuracy by 6 to 13%. Also removing "dead" words from the lexicon had a good impact on the function of the ASR system, because the number of words in the lexicon was decreased by 7 to 12% simultaneously with maintaining or small increase in the recognition accuracy.

Among all recognition results you can notice a relatively low level of the recognition accuracy for the Tests 3 compared with the other outcomes. We investigated the corresponding speech data and found out a very surprising fact – owing to the break-down of the acoustic channel, the most part of the TV commentary was transmitted through a telephone channel using a phone set! It is evident that speech transmitted over a telephone channel with its narrow frequency band can not be recognized so well using acoustic models prepared for quite different conditions.

5. CONCLUSIONS

The outcomes of the baseline ASR system designed for transcriptions of TV ice-hockey commentaries indicate the feasibility of this task. To improve the function of such a system we will have to solve problems with the noise in the future. In the above experiments we, perhaps competently, supposed that the acoustic models, trained using a large portion (more than 25 hours) of a speech immersed in a various and often very strong noise, cover well all ice-hockey backgrounds. Also the decrease in OOV words by collecting and annotating more data should bring further improvement of the ASR system.

6. REFERENCES

[1] Byrne, W., Hajič, J., Ircing, P., Jelinek, F., Khudanpur, S., Krbec, P., and Psutka, J.: On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language – Czech. –In: Eurospeech'2001, Aalborg, 2001, pp.487-490.

[2] Stolcke, A.: SRILM – The SRI Language Modeling Toolkit.

[3] Mohri, M., Riley, M., Pereira, F.: Weighted Finite-State Transducers in Speech Recognition. –In: Intern. Workshop on Automatic Speech Recognition 2000.